How to Measure the Agreement between Sequences: A Comment

Dijkstra, W.

published in
Sociological Methods and Research
2001

DOI (link to publisher)
10.1177/0049124101029004006

document version
Publisher's PDF, also known as Version of record

How to Measure the Agreement between Sequences : A Comment

WIL DIJKSTRA

Sociological Methods & Research 2001 29: 532
DOI: 10.1177/0049124101029004006

Some problems of optimal alignment procedures to measure the agreement between sequences are discussed. Hidden Markov models may be a new approach that is especially suited for grouping sequences that are more or less alike.

How to Measure the Agreement Between Sequences

A Comment

WIL DIJKSTRA
Vrije Universiteit

The SEQUENCE program (Dijkstra 1994, 1999) offers a large number of methods for studying sequences, like lag-sequential analysis, tree analysis, and sophisticated procedures to find particular sequential patterns in large sets of sequences. It also includes the classical optimal alignment procedure. *Optimal alignment* is based on the number of operations that is necessary to make one sequence equal to another sequence. Consider, for example, the sequences AC and ABC, then one may insert B in the first sequence (or delete B from the second sequence) to make both sequences equal. Or to make ABC and ABD equal, one may substitute C with D in the first sequence. Generally, the less similarity there is between the sequences, the more operations (insertions and deletions, or "indels," and substitutions) are necessary. If the sequences are completely different (e.g., ABC and DEF), one needs the maximum number of operations (e.g., three substitutions). The problem in optimal alignment is finding the least number of operations that are necessary to turn one sequence into the other. By comparing this number with the maximum number of operations, given the length of the sequences, one has a measure of (dis)agreement between 0 and 1. This is done by recursive programming. In addition, one may weight indel and substitution operations differently.

SOCIOLOGICAL METHODS & RESEARCH, Vol. 29 No. 4, May 2001 532-535

Despite the elegance of the optimal alignment solution, there are two problems with it. First, the agreement measure it yields is sometimes somewhat unrealistic. Consider the sequences XYAB and ABPR. To make both sequences alike, one needs four operations, for example, deleting X and Y from the first sequence and P and R from the second one. Or one may substitute X by A, Y by B, A by P and B by R. Hence, one should conclude that both sequences are completely dissimilar. Nevertheless, it seems reasonable that the sequence pair XYAB-ABPR is more similar than, for example, XYAB-DEFG. The problem can be solved by giving substitutions more weight than indel operations. This introduces a different problem, however, namely, what weight should be assigned.

A second drawback of the optimal alignment solution is the speed of the calculations. If one wants to perform a cluster analysis on a large set of sequences, one has to calculate the agreement between each pair of sequences in the data set. We are working now with data sets of more than 7,000 sequences, amounting to about 25 million sequence pairs.

Therefore, our problem was to develop an agreement measure that did not have the above-mentioned problem of (sometimes) unrealistic measures, without having to decide on arbitrarily weighting substitution costs, and that could be calculated much faster. The agreement measure we developed, as described in Dijkstra and Taris (1995), fulfils these requirements: It does not have the problem of unrealistic measures, it is not necessary to specify weights for particular operations, and it is about 100 times as fast as the optimal alignment procedure.

In our SEQUENCE program, there are two options for calculating our agreement measure (DT for short), an exact solution and an approximate solution. The approximate solution is about twice as fast as the exact solution. The approximate solution nearly always gives exactly the same agreement measure as the exact solution. For example, in the data set of 494 sequences we used in our example in Dijkstra and Taris (1995), there were 121,771 pairs of sequences; for all pairs, the exact and approximate solution yielded exactly the same result. The difference between the exact and the approximate solution concerns the point raised by the anonymous reviewer of the Van Driel and Oosterveld (2001 [this issue]) article. To make the sequence

AADAAG equal to the sequence GAD, three superfluous codes "A" have to be removed from AADAAG. In the approximate solution, the program makes a "best guess" about which As should be removed. In the exact solution, those As are removed to reach the optimal solution.

The criticism of Van Driel and Oosterveld concerns the approximate solution. If they would have applied the exact solution, they would have discovered that the program indeed gives the optimal solution. Both the program and the manual clearly state that the approximate algorithm may give a nonoptimal solution; the approximate solution is incorporated only for purposes of speed. A demo version of the program and the manual can be downloaded from our Web site at http://svn.scw.vu.nl/sequence/. I have to insist that their claim that DT gives a nonoptimal solution is simply not true: The exact algorithm does give the optimal solution. Nevertheless, I have to thank the authors for drawing the attention to our SEQUENCE program.

A main reason to apply agreement measures to sets of sequences is to group sequences in a meaningful way. Let me finish my comment by mentioning an exciting different approach to this problem, the hidden Markov model (HMM), widely used for pattern recognition like speech recognition and handwriting recognition (Bengio 1999). In a first-order Markov process, the probability of the occurrence of an event can be predicted using only the information from the preceding event. Such an assumption is too strong to hold in most social science research. In an HMM, however, it is assumed that the sequence of observations is the output of a hidden (unobservable) state sequence. The problem, then, is to find the "hidden" sequence that "explains" an observed set of sequences and next to find a small number of hidden sequences, each "explaining" a group or cluster of sequences (see, e.g., Eddy 1998; Krogh 1998; and Smyth 1997). This approach seems to be very promising, but here, too, successful application of these procedures to large data sets calls for much faster computers.

REFERENCES

Bengio, Yoshua. 1999. "Markovian Models for Sequential Data." *Neural Computing Surveys* 2:129-62.

Dijkstra, Wil. 1994. "SEQUENCE—A Program for Analysing Sequential Data." *Bulletin de Méthodologie Sociologique* 43:134-42.

———. 1999. "A New Method for Studying Verbal Interactions in Survey-Interviews." *Journal of Official Statistics* 15:67-85.

Dijkstra, Wil and Toon Taris. 1995. "Measuring the Agreement Between Sequences." *Sociological Methods & Research* 24 (2): 214-31.

Eddy, Sean R. 1998. "Profile Hidden Markov Models." *Bioinformatics* 14:755-63.

Krogh, Anders. 1998. "An Introduction to Hidden Markov Models for Biological Sequences." Pp. 45-63 in *Computational Methods in Molecular Biology*, edited by Steven L. Salzberg, David B. Searls, and Simon Kasif. Amsterdam, the Netherlands: Elsevier.

Smyth, Pahdraic. 1997. "Clustering Sequences With Hidden Markov Models." Pp. 648-54 in *Advances in Neural Information Processing* Vol. 9, edited by Michael C. Mozer, Michael I. Jordan, and Thomas Petsche. Cambridge, MA: MIT Press.

Van Driel, Kees, and Paul Oosterveld. 2001. "Nonoptimal Alignment: A Comment on 'Measuring the Agreement Between Sequences' by Dijkstra and Taris." *Sociological Methods & Research* 29:524-31.

Wil Dijkstra is an associate professor in the Department of Social Research Methodology at Vrije Universiteit, Amsterdam. His primary interest is in studying the course of actions of interviewer and respondent in survey interviews (question-answer sequences). His SEQUENCE program ended as finalist in the European Academic Software Award 2000. He has written numerous articles and several books about survey research methods, most recently "A New Method for Studying Verbal Interactions in Survey-Interviews" (1999) in Journal of Official Statistics *and (with J. H. Smit) "Persuading Reluctant Recipients in Telephone Surveys," forthcoming in* Survey Nonresponse, *edited by R. M. Groves, D. A. Dillman, J. L. Eltinge, and R.J.A. Little.*